A Literature Study on Human Motion Analysis Using Depth Imagery

SANDHYA PAI
M.Tech-Student
Computer Science and Engineering,
R V College of Engineering,
Bangalore, Karnataka, India

Abstract: **Analysis of human behavior through visual information is a highly active research topic in the computer vision community. This analysis is achieved in the literature via images from the conventional cameras; however recently depth sensors are used to obtain new type of images known as depth images. This human motion analysis can be widely applied to various domains, such as security surveillance in public spaces, shopping centers and airports. Home care for elderly people and children can use live video streaming from an integrated home monitoring system to prompt timely assistance. Moreover, automatic human motion analysis can be used in Human–Computer/Robot Interaction (HCI/HRI), video retrieval, virtual reality, computer gaming and many other fields. Human motion analysis using a depth sensor is still a new research area. Most work is focused on motion capture of articulated body skeletons. However, the research community is showing interest in higher level action related research. This report explains the advantages of depth imagery and then describes the new categories of depth sensors such as Microsoft Kinect that are available to obtain depth images. High-resolution real-time depth images are cheaply available because of tools like Microsoft Kinect. The main published research on the use of depth imagery for analyzing human activity is reviewed. A growing research area is the recognition of human actions and hence the existing work focuses mainly on body part detection and pose estimation. The publicly available datasets that include depth imagery are listed in this report, and also the software libraries that are available for the depth sensors are explained. With the development of depth sensors, an increasing number of algorithms have employed depth data in vision-based human action recognition. The increasing availability of depth sensors is broadening the scope for future research. This reports provides an overview of this emerging field followed by various vision based algorithms used for human motion analysis.**

I. INTRODUCTION TO HUMAN MOTION ANALYSIS

Human motion analysis has been a highly active research area in computer visions, whose goal is to automatically segment, capture and recognize human motion in real time, and perhaps predict ongoing human activities. Home care for elderly people and children could use live video streaming from an integrated home monitoring system to prompt timely assistance. If machines could automatically interpret the activities people perform in everyday life, many tasks would be revolutionized. For example, automatic human motion analysis can be used in Human–Computer/Robot Interaction (HCI/HRI), video retrieval, virtual reality, computer gaming and many other fields. It can also be applied to various domains, such as security surveillance in public spaces, including shopping centers and airports

II. RESEARCH LITERATURE STUDY

Various systems use depth information this way to extract the scene foreground (Grammalidis et al., 2001[13]; Schwarz et al., 2011[8], 2012[23]; Van and Van (2011) [1]; Schwarz et al. (2012) [2]; Jansen et al. (2007) [3]; Guomundsson et al. (2008) [4]) built a freehand interactive surface system called DepthTouch to track hand gestures using a ZSense depth camera and chose a simple 3D region of interest to segment a user's hands. There has not been much published work on the issue of holes in depth images. This is because until recently most work used a ToF camera. Zhang and Parker (2011) [5] use Kinect, and mention a hole filling process performed along with noise reduction, although they do not give precise details. To model a 3D human body, first the body parts must be found. These may include the head, torso, arms, hands, legs and feet. Several methods have been presented for finding human body parts from depth imagery (Plagemann et al., 2010 [6]; Siddiqui and Medioni, 2010 [7]; Kalogerakis et al., 2010[10]; Holt et al., 2011[14]; Shotton et al., 2012 [12]; Anguelov et al. (2005) [9]; Zhu and Fujimura (2007) [11]; Schwarz et al. (2011) [8]). Space-time approaches treat each action video as a 3D volume along spatial (x; y) and temporal (t) axes. The video can be processed either as a whole (Holte and Moeslund, 2007 [24]; Roh et al., 2010 [25]; Ni et al., 2011 [26]; Wu et al., 2012 [27]), or as collection of local feature points (Li et al., 2010 [29]; Zhang and Parker, 2011 [5]; Ni et al., 2011[26]; Malgireddy et al., 2012[30]). Generally, these approaches are suitable for simple actions such as walking, running, jumping and waving. Depth images viewed as a function of time (Wang et al. (2012) [31]; Jansen et al. (2007) [3]; Chen et al. (2011) [32]; Reyes et al. (2011) [33]; Sempena et al., 2011 [34]; Sung et al. (2012) [35]; Xia et al. (2012) [36]; Wang et al. (2012) [37]; Shotton et al. (2012) [12]) can be used to analyse the human motion. There are three types of sensors discussed in this section namely stereo cameras, ToF cameras and Structured light sensors. Apart from these there are many sensors that can capture range data. For instance, a 3D body scanner from Cyberware and the Minolta 3D scanner which has been used for the face recognition.

Stereo Cameras

Stereo machine vision is biomimetic: it is inspired by human vision. Stereo cameras have been made into products especially for research use, such as the Bumblebee series from Point Grey Research. It infers the 3D structure of a scene from two (or more) images from different viewpoints, as is the case for human stereo vision. Depth map acquisition from stereo vision is an important computer vision research field dating back to the 1960s. Stereo intensity images are sensitive to light changes, which increases difficulties with correspondence matching for triangulation. Due to the complexities of stereo geometry calculation, reconstruction of a depth map from stereo images still remains a challenge. These issues make the depth map reconstruction from stereo vision still impractical for real-time real-world applications.

Time-of-Flight Cameras

Human stereo vision works well, however this does not mean machines must view the world as humans do. To make a robust method, different sensing technologies may be adopted. In contrast with stereo vision, a time-of-flight (ToF) camera estimates distance to an object surface using active light pulses from a single camera, whose time to reflect from the object, together with the speed of light, give the distance. Compared with other laser 3D scanning devices, ToF cameras are cheaper and smaller. Most current commercial devices use a sinusoidally modulated infrared light signal, and distance is estimated using the phase shift of the reflected signal on a standard CMOS or CCD detector. The resolution of the depth image is currently between 64 48 and 200 200 pixels, and the range varies from 5m to 10m. Because their distance calculation is computationally simple, ToF cameras can achieve high frame rates, which make them suitable for real-time applications. The main advantages of ToF cameras are their high speed, and their dense depth map that covers every pixel. The major practical drawback is their high price although they are still cheaper than some other 3D scanning devices. The major technical drawback is the low resolution.

Structured Light Sensors

Microsoft released an imaging device called Kinect, 4 which is priced at a consumer level for domestic use. Kinect consists of an RGB camera and a depth sensor. The depth sensor provides images at640 480 pixels and 30 frames per second. The range is around 0.8– 3.5 m, with a resolution of about 1 cm. Kinect computes depth from structured light, which is a topic that has been studied since the 1970s. The idea is based on stereo vision. One camera is replaced by a light source that projects a known pattern, hence the light is structured. The Kinect depth sensor consists of an infrared projector and infrared CMOS sensor. An irregular pattern of dots is projected onto the scene, and the depth measurement is based on triangulation. The advantage of structured light devices over ToF devices is that they are much cheaper. This makes them suitable for everyday applications. A major issue with structured light is that the depth images have holes because some areas cannot be seen by both the projector and the camera. ToF cameras do not have this problem.

Sensing Summary and Comparison

The ranges of ToF and structuredlight sensors are limited by the distance to which the light can penetrate and be reflected back because they are active vision systems. The range of stereo cameras is only limited by how the baseline is set, and the ambient light in the scene. Stereo cameras and structured light have holes in the depth images because some locations are not visible to both cameras. Because ToF cameras have a single viewpoint there will be no holes in the images. Table compares three types of depth sensors. Apart from these three types of sensors, there are other sensors that can capture range data. For instance, a 3D body scanner from Cyberware. In recent years, large advances have made depth cameras cheap and readily available for research and domestic use. This has caused a change in the research community, which is now developing new directions of research on imagery from ToF cameras and Kinect. In particular, Kinect has opened new possibilities in human motion analysis.

Table - Comparison of Depth Sensors

Sensor type	**Stereo cameras**	**Time of flight**	**Structured light**
Resolution	High: 640 480 or more	Low: 64 48 to 200 200	High: 640 480
Speed	Slow	Fast	Fast
Range	Only limited by baseline	Varies from 5 m to 10 m (indoors	0.8 – 3.5 m (typically indoors)

		or outdoors)	
Depth resolution	Depends on camera baseline and resolution	Less than 5 mm	Less than 1 cm
Field of views	Not limited; Depends on camera lenses	Approx. 43L(v), 69L(h)	43L(v), 57L(h)
Holes in depth map	Yes	No	Yes
Price	Cheap	Expensive	Cheap
Sensitive to lighting	Yes	No	No

Body Part Detection

Siddiqui and Medioni (2010) [7] built different detectors for head, forearm and hand with kinematic constraints. Segmentation and labelling of objects from 2D or 3D data is an important research area in computer vision. Most work does not aim especially at segmenting human body parts, however a few projects have demonstrated articulated human body segmentation. Anguelov et al. (2005) [9] used Markov Random Fields (MRFs) to segment articulated wooden puppets into head, limbs, torso and background, from a set of depth images. They apply their approach to many types of objects, not just human bodies. Similarly Kalogerakis et al. (2010) [10] used a datadriven Conditional Random Field (CRF) model to segment and label object parts from a closed 3D mesh. They showed that a human body mesh can be segmented into eight parts. Zhu and Fujimura (2007) [11] use handengineered heuristics for coarse upper body part labelling based on depth constraints, color constraints, and coherence between successive frames. This is used as the first step of upper body pose estimation.

Body Pose Modelling

Some early work with depth images fitted a body model directly to the image. Grammalidis et al. (2001) [13] proposed an iterative approach to estimate MPEG-4 Body Animation Parameters (BAPs) of an arm by minimizing the error between synthetic and real depth images. Downhill simplex minimization was used for iterative minimization, which is sensitive to local minima, and requires a good initial guess of the BAPs. Shotton et al. (2012) [12] proposed a skeleton model where the joints are fitted to previously labelled body parts using mean shift. Their system is based on depth pixels rather than converting to a 3D point cloud. Holt et al. (2011) [14] proposed an approach for static upper body pose estimation through 'poselet' detection and classification, which does not track the pose. They also used a randomized decision forest for the classification task. They claim their approach does not require a large amount of training data. Charlesand Everingham (2011) [15] inferred an articulated 2D human pose from a body silhouette extracted from a single depth image using a Pictorial Structure Model (PSM). The main contribution is that, instead of a conventional rectangular limb model, they model each limb with a mixture of probabilistic shape templates, which showed a promising improvement accuracy.

Zhu and Fujimura (2007) [11] proposed a method for human upperbody (hand, torso and arms) pose estimation and tracking from ToF depth sequences. They first label the upper body parts then fit a 3D body model by inverse kinematics constraints using ICP on each body part. A T-pose initialization is required to scale the skeleton model. ICP is a popular algorithm for fitting a model to a 3D point cloud, but a major drawback is the need for a good initial position and it is difficult to recover from a tracking failure. Other methods have also been used. Siddiqui and Medioni (2010) [7] model a 3D body with fixedwidth cylinders based on relative distance and anglesbetween body parts. The main contribution of their work is applying a datadriven Markov Chain Monte Carlo (MCMC) approach, comparing it with an ICP-based approach, and demonstrating their MCMC approach outperforms ICP. Their method does not require a large training dataset, however, it is sensitive to fast motion and occlusions. Pellegrini and Iocchi (2008) [19] built a 3D body model composed of two sections: a head-torso block and a leg block with three principle joints: head, pelvis, and the legs' contact with the floor. Their model does not contain arms. Pose is estimated using the five angles between different parts. This system is limited to observation of very simple human movement without details, from sources like surveillance imagery. Zhu et al. (2008) [20] proposed another method based on their previous work (Zhu and Fujimura, 2007 [11]). They built a key point detector to define anatomical landmarks for a human upper body. This also requires an open-arm pose for initialization. Zhu and Fujimura (2010) [21] further extend their work to full-body motion tracking. They addressed robustness of continually tracking body parts through self-occlusion, which sometimes caused failure of their previous method. Schwarz et al. (2010, 2012) [22] [23] estimate a coarse full-body model from a 3D point cloud. They first find the centroid of the body and the external bounding points. Poses are represented simply by a list of the vectors between these points. Accuracy comes mostly from the priors on body postures.

Space-time Approaches

Some previous work on intensity images has successfully represented an action as a 3D shape in space-time. A template matching method can be applied to find the nearest action. An example of this is the movement of a single person's silhouette stacked over time, as in the Motion History Image (MHI). The recognition is done by estimating the similarity of the captured volume to previously labelled volumes. Such approaches may require human body shape extraction, such as a body silhouette. In intensity images this is still a challenging task, but in depth images foreground can be easily extracted. Holte and Moeslund (2007) [24] proposed a basic approach to recognize one and two-arm gestures. They use double difference range images to detect the movements. Each arm gesture is modelled using shape contexts based on a spherical histogram centered on the upper body. For each gesture, start and end points must be given. *(Ni et al., 2011) [26] extended MHIs to create 3D-MHIs by adding two more channels: forward motion history (fDMHI) and backward motion history (bDMHI) calculated by thresholding depth changes. They tested the approach on their dataset (HuDaAct) and results showed an improvement of nearly 30% recognition accuracy compared with the original MHI approach. Wu et al. (2012) [27] also extended MHIs (Extended-MHI) by combining two more elements: gait energy information (GEI) and inversed recording (INV), which are designed to handle the poor performance of the original MHIs at representing repetitive actions and to recover the loss of the initial frames' action information, respectively. The 3D depthinformation is not used explicitly, but only as an intensity image would be. Ni et al. (2011) [26] evaluated the extension of 2D spatio-temporal features to 3D space by adding depth information for action recognition. They simply divide the depth range into multiple bins, and create a code word histogram for each bin using Histogram of Orientated Gradient (HOG)/Histogram of Optical Flow (HOF). Obviously, their method is not invariant to translation along the z axis, however, it obtained good results for the dataset on which it was tested, where each action occurs in the same spatial region. They compared performance with a conventional approach using spatio-temporal bag of local features, and the results showed a small improvement in recognition accuracy.*

Tracking Based Approaches

Current space-time approaches can recognize simple human actions by image appearance, but they cannot handle complex activities. Since the emergence of depth sensors, 3D human body part tracking has become feasible for highlevel recognition tasks. In particular, the middleware for Kinect has provided robust human skeleton tracking. Recently, several approaches have built on this. The advantage is that more complex human actions can be modelled, higherlevel algorithms can be based on the skeleton data. Actions may involve interactions with other humans or objects, although this may rely heavily on robust object labelling, detection or tracking. Algorithms such as layered Hidden Markov Models (HMMs), Convolutional Neural Networks (CNNs), Conditional Random Fields (CRFs), Allen's Interval Algebra (IA), Probabilistic Petri Nets (PPN), and Dynamic Time Warping (DTW) can be applied for tracking based systems. The scene captured from a depth camera can be combined with known 3D positions in the environment. For long term indoor monitoring the environment can be defined, then the 3D position of a person used in activity recognition. Jansen et al. (2007) [3] use a simple distance constraint on the height of a person's silhouette to recognize if the person is standing, sitting or lying. They claim the system is useful for elderly people's home care. This is early work with very basic use of depth constraints to detect a person's state. Chen et al. (2011) [32] address home monitoring using depth sensors, by aiming to recognize domestic activities such as drinking. Their approach uses distance between body parts and objects over time, and models each activity via spatio-temporal reasoning using Allen's Interval Algebra (IA). This approach could be extended to recognize higher level activities by adding more complex algebraic expressions. Reyes et al. (2011) [33] represent a human model using a feature vector defined by 15 joints on a 3D human skeleton model. The model is obtained using the Primesense human skeleton API. They use Dynamic Time Warping (DTW) with automatic feature weighing on each joint to achieve real-time action recognition. Similarly, (Sempena et al., 2011 [34]) also used the Prime Sense API with DTW. The skeleton joints are represented using quaternions to form a 60element feature vector for 15 joints in total. Sung et al. (2012) [35] proposed a twolayered Maximum Entropy Markov Model (MEMM) to recognize domestic single person activity. The activity is modelled in each frame using a 459element feature vector from various body joints obtained from the Prime sense API. They claim their hierarchical MEMM has an advantage because a single state may connect to different parents for only a specified period of time, which would not be feasible in a Hierarchical Hidden Markov Model (HHMM). They tested on twelve activities and achieved an average recognition accuracy of 64.2%. Xia et al. (2012) [36] proposed a histogram technique on 3D joint locations (HOJ3D) using modified spherical coordinates. HMMs are applied for the classification task. The main advantage is real-time performance. Wang et al. (2012) [37] obtained and tracked 20 body joints using the method from Shotton et al.

(2012) [12]. They used Fourier Temporal Pyramid (FTP) to model the temporal patterns of the joint feature vectors. Their main contribution is an Actionlet Ensemble (AE) model that can handle errors of the skeleton tracking and better characterize the intra-class variations. Theytested on the MSR-Action3D dataset (Li et al., 2010 [29]), achieving 88.2% recognition accuracy.

III. CONCLUSIONS

With the development of depth sensors, and especially the emergence of Microsoft Kinect, an increasing number of algorithms have employed depth data in vision based human action recognition. Computer vision researchers are exploring an extended research field with many potential applications. Preprocessing for depth images is described, as it has not been adequately addressed before. The comprehensive review addresses articulated 3D body modelling for human pose estimation and human action recognition. Datasets and open libraries used for development of these algorithms are listed. There has been much research in building algorithms on intensity imagery. Depth imagery may be processed with the same algorithms, as in some of the previous work (such as local feature detection), but would ideally have modified or new algorithms to suit its particular properties. These algorithms will be developed over time, as will techniques to combine intensity and depth imagery, using their advantages to complement each other and achieve better and more robust solutions. One big challenge for improving human action recognition is the lack of large and realistic action datasets, with wide ranges of human body shape, diversity of body movements and ground truth labels. Significant work has already been conducted on pose estimation from depth data. Higher resolution body part modelling including finger details still needs further exploration, to enable subtle hand gesture recognition and interaction tasks. A promising direction for future work is development of more sophisticated highlevel activity recognition, which should allow processing of interaction with objects and other people, and group activity. The most appropriate machine learning techniques should be chosen to allow this new depth imagery to be fully exploited in under-standing human behavior.

IV. ACKNOWLEDGEMENTS

The author thanks Dr.S.Sridhar, Professor and Director, Cognitive & Central Computing, R.V.College of Engineering, Bangalore, India for communicating this article to this Journal for publication after modifications. The author would like to thank the guides, Dr.Roopalakshmi R, Associate Professor and Swarnalatha K S, Asst.Professor, Department of Computer Science and Engineering, RVCE for their help, sharing their technical expertise and timely advice and also thank Dr.Shobha G, HoD of CSE Department.

V. REFERENCES

[1]. Van den Bergh, M., Van Gool, L., “Combining RGB and ToF cameras for real-time 3D hand gesture interaction”, In proc of IEEE Workshop on Applications of Computer Vision (WACV),pp. 66–72,2011

[2]. Schwarz, L.A., Mkhitaryan, A., Mateus, D., Navab, N., “Human skeleton tracking from depth data using geodesic distances and optical flow”, Image Vision Comput. 30, 217–226, 2012.

[3]. Jansen, B., Temmermans, F., Deklerck, R., “3D human pose recognition for home monitoring of elderly”. In proc of: Engineering in Medicine and Biology Society, 29th Annual Internat. Conf. of the IEEE (EMBS), pp. 4049–4051, 2007.

[4]. Guomundsson, S., Larsen, R., Aanaes, H., Pardas, M., Casas, J., “TOF imaging in smart room environments towards improved people tracking”,In proc of: IEEE Comput. Society Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1–6., 2008

[5]. Zhang, H., Parker, L.E., “4-Dimensional local spatio-temporal features for human activity recognition”,In proc of: 2011 IEEE/RSJ Internat. Conf. on Intelligent Robots and Systems (IROS),pp. 2044–2049, 2011

[6]. Plagemann, C., Ganapathi, V., Koller, D., Thrun, S, “Real-time identification and localization of body parts from depth images”,In proc of: IEEE Internat. Conf. on Robotics and Automation (ICRA), 2010.

[7]. Siddiqui, M., Medioni, G., “Human pose estimation from a single view point, real-time range sensor”,In proc of: 2010 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW),pp. 1–8, 2010.

[8]. Schwarz, L.A., Mkhitaryan, A., Mateus, D., Navab, N., “Estimating human 3D pose from time-of-flight images based on geodesic distances and optical flow”, In: FG,pp. 700–706, 2011.

[9]. Anguelov, D., Taskarf, B., Chatalbashev, V., Koller, D., Gupta, D., Heitz, G., Ng, A., “Discriminative learning of Markov random fields for segmentation of 3Dscan data”,In proc of: IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR),pp. 169–176,2005.

[10]. Kalogerakis, E., Hertzmann, A., Singh, K., “Learning 3D mesh segmentation and labeling”, In: ACM SIGGRAPH 2010, ACM, New York, NY, 775 USA, pp. 102:1–102:12, 2010.

[11]. Zhu, Y., Fujimura, K., “Constrained optimization for human pose estimation from depth sequences”, In proc. of the Eighth Asian Conf. on Computer Vision, Part I, Springer-Verlag, Berlin, Heidelberg,pp. 408–418, 2007.

[12]. Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., Blake, A., “Efficient human pose estimation from single depth images”, IEEE Trans. Pattern Anal. Machine Intell, 2012.

[13]. Grammalidis, N., Goussis, G., Troufakos, G., Strintzis, M., “3-D human body tracking from depth images using analysis by synthesis”, In proc. of 2001 Internat. Conf. on Image Processing, vol. 2, pp. 185–188, 2001.

[14]. Holt, B., Ong, E.J., Cooper, H., Bowden, R., “Putting the pieces together: Connected poselets for human pose estimation”,In proc of: 2011 IEEE Internat. Conf. on Computer Vision Workshops (ICCV Workshops), pp. 1196–1201, 2011.

[15]. Charles, J., Everingham, M., “Learning shape models for monocular human pose estimation from the Microsoft Xbox Kinect”, In proc of: 2011 IEEE Internat. Conf. on Computer Vision Workshops (ICCVW), pp. 1202–1208, 2011.

[16]. Grest, D., Woetzel, J., Koch, R., “Nonlinear body pose estimation from depth images”, In proc. of the 27th DAGM Conf. on Pattern Recognition. Springer-Verlag, Berlin, Heidelberg,pp. 285–292, 2005.

[17]. Demirdjian, D., Ko, T., Darrell, T., “Constraining human body tracking”, In proc. of Ninth IEEE Internat. Conf. on Computer Vision, vol. 2, pp. 1071–1078, 2003.

[18]. Ganapathi, V., Plagemann, C., Koller, D., Thrun, S., “Real time motion capture using a single time-of-flight camera”, In proc of: 2010 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 755–762, 2010.

[19]. Pellegrini, S., Iocchi, L., “Human posture tracking and classification through stereo vision and 3D model matching”, J. Image Video Process, 7:1–7:12, 2008.

[20]. Zhu, Y., Dariush, B., Fujimura, K., “Controlled human pose estimation from depth image streams”, In proc of: IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW),pp. 1–8, 2008.

[21]. Zhu, Y., Fujimura, K., “A bayesian framework for human body pose tracking from depth image sequences”, Sensors 10, 5280–5293, 2010.

[22]. Schwarz, L.A., Mateus, D., Castaneda, V., Nava, N., “Manifold learning for tof-based human body tracking and activity recognition”, In proc of: British Machine Vision Conf. (BMVC), 2010.

[23]. Schwarz, L.A., Mateus, D., Navab, N., “Recognizing multiple human activities and tracking full-body pose in unconstrained environments”, Pattern Recognit. 45, 11–23, 2012.

[24]. Holte, M.B., Moeslund, T.B., “Gesture recognition using a range camera. Technical Report”,pp. 1–5, 2007.

[25]. Roh, M.C., Shin, H.K., Lee, S.W., “View-independent human action recognition with volume motion template on single stereo camera”, Pattern Recognition Lett. 31, 639–647, 2010.

[26]. Ni, B., Wang, G., Moulin, P., “RGBD-HuDaAct: A color-depth video database for human daily activity recognition”,In proc of: 2011 IEEE Internat. Conf. on Computer Vision Workshops (ICCV Workshops), pp. 1147–1153, 2011.

[27]. Wu, D., Zhu, F., Shao, L., “One shot learning gesture recognition from RGBD images”, In proc of: 2012 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW),pp. 7–12, 2012.

[28]. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S., “Behavior recognition via sparse spatio-temporal features”, In proc of: Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS), 65–72, 2005.

[29]. Li, W., Zhang, Z., Liu, Z., “Action recognition based on a bag of 3D points”, In proc of: 2010 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW),pp. 9–14, 2010.

[30]. Malgireddy, M., Inwogu, I., Govindaraju, V., “A temporal bayesian model for classifying, detecting and localizing activities in video sequences”, In proc of : 2012 IEEE Computer Society Conf. on Computer Vision

and Pattern Recognition Workshops (CVPRW), pp. 43–48, 2012.

[31]. Wang, J., Liu, Z., Chorowski, J., Chen, Z., Wu, Y., “Robust 3D actionrecognition with random occupancy patterns”, In: Computer Vision – ECCV 2012, pp. 872– 885, 2012.

[32]. Chen, L., Wei, H., Ferryman, J., “Recognition of everyday domestic activities using a depth sensor”, In: BMVC 2011 Student, Workshop,pp. 27–37, 2012.

[33]. Reyes, M., Dominguez, G., Escalera, S., “Featureweighting in dynamic timewarping for gesture recognition in depth data”, In proc of: 2011 IEEE Internat. Conference on Computer Vision Workshops (ICCVW), pp. 1182–1188, 2011.

[34]. Sempena, S., Maulidevi, N., Aryan, P., “Human action recognition using dynamic time warping”, In proc of: 2011 Internat. Conf. on Electrical Engineering and Informatics (ICEEI), pp. 1–5, 2011.

[35]. Sung, J., Ponce, C., Selman, B., Saxena, A., “Unstructured human activity detection from RGBD images”, In proc of: 2012 IEEE Internat. Conf. on Robotics and Automation, 2012.

[36]. Xia, L., Chen, C.C., Aggarwal, J., “View invariant human action recognition using histograms of 3D joints”, In proc of: IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW),pp. 20–27, 2012.

[37]. Wang, J., Liu, Z., Wu, Y., Yuan, J., “Mining actionlet ensemble for action recognition with depth cameras”, In proc of: 2012 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1290–1297, 2012.

[38]. Wolf, C., Mille, J., Lombardi, L., Celiktutan, O., Jiu, M., Baccouche, M., Dellandrea, E., Bichot, C.E., Garcia, C., Sankur, B., “The LIRIS Human activities dataset and the ICPR 2012 human activities recognition and localization competition”, Technical Report, RR-LIRIS-2012-004, LIRIS Laboratory, 2012.

[39]. Lulu Chen, Hong Wei, James Ferryman, “A survey of human motion analysis using depth imagery”, Pattern Recognition Letters 34, 1995-2006, 2013.